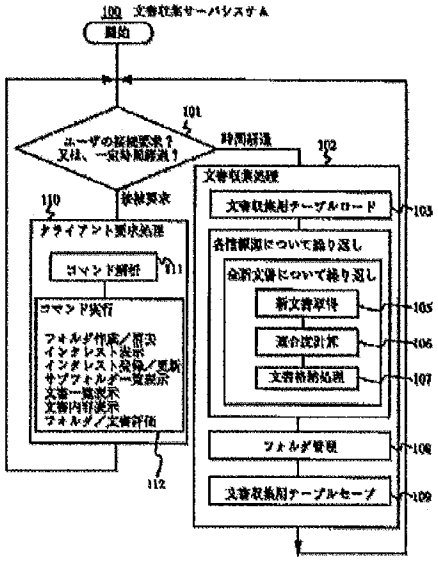
DOCUMENT INFORMATION CLASSIFYING METHOD, AND METHOD AND SYSTEM FOR DOCUMENT INFORMATION COLLECTION USING THE SAME

Publication number: JP7049875
Publication date: 1995-02-21
Inventor: YUASA HIROKO; KOJIMA KEIJI
Applicant: HITACHI LTD
Classification:
- international: ***G06F12/00; G06F17/21; G06F17/27; G06F17/30;*** **G06F12/00; G06F17/21; G06F17/27; G06F17/30;** (IPC1-7): G06F17/30; G06F12/00; G06F17/27
- European:
Application number: JP19930195839 19930806
Priority number(s): JP19930195839 19930806

Abstract of **JP7049875**

PURPOSE:To determine a corresponding folder and obtain proper information classifications by classifying information in consideration of the degree of matching between retrieval conditions which are stored corresponding to respective folders and retrieved document information and the hierarchical structure of the retrieval conditions. CONSTITUTION:A document collecting server system 100 accesses external information sources periodically and starts a document collecting process 102. When a table for document collecting is loaded, documents stored in the respective information sources are acquired and documents meeting the retrieval condition group that the user has registered are retrieved. At this time, the frequency of appearance of each word in the retrieval conditions in an object document is decided as the degree of matching between the object document and retrieval conditions through matching degree calculation 106. Then a document storing process 107 selects a folder where the object document is classified in consideration of the hierarchical structure of the folders among folders where the matching retrieval conditions are registered and stores it inside. Further, a folder managing process 108 rearranges documents corresponding to the gathering state of documents by automatically dividing the folder in which many documents are stored.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平7－49875

(43)公開日　平成７年(1995)２月21日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/30					
	12/00	５２０　Ａ	8944－5B			
	17/27					
			9194－5L	Ｇ０６Ｆ　15/ 401	３１０　Ｄ	
			7315－5L	15/ 20	５５０　Ｆ	

審査請求　未請求　請求項の数28　ＯＬ　（全 19 頁）

(21)出願番号　特願平5－195839

(22)出願日　平成５年(1993)８月６日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　湯浅　寛子
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　小島　啓二
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　弁理士　小川　勝男

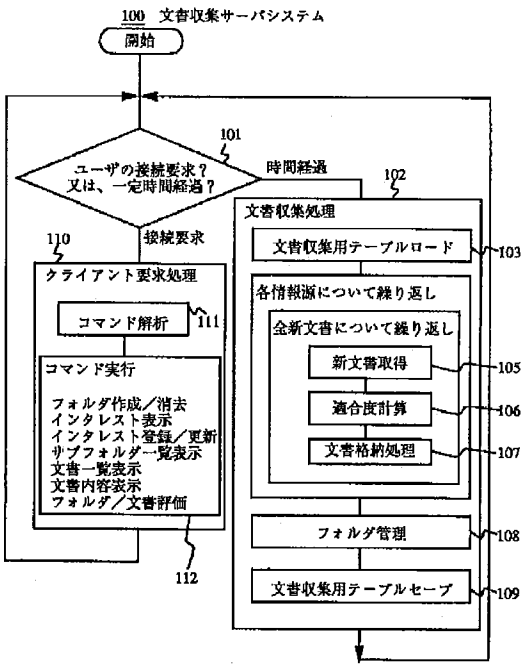
(54)【発明の名称】　文書情報分類方法およびそれを用いた文書情報収集方法、文書情報収集システム

(57)【要約】　（修正有）

【構成】　文書収集サーバシステム１００は、自動的に複数の情報源に接続して新文書を取得し、適合度計算１０６によって、あらかじめユーザが記述した検索条件との適合度を調べる。文書格納処理１０７は、検索条件間の関係から分類体系を構成し、適合した文書を分類してフォルダに格納する。フォルダ管理処理１０８は、各フォルダへの情報の集まり具合を監視し、自動的にフォルダの細分化、統合、構造の変更を行なって情報の整理をする。

【効果】　各分類への情報の集まり具合に応じて、分類体系や検索条件を改善し、各分類に分類される情報量をその全体を容易に把握できる程度の数に抑さえることができる。

図２



1

【特許請求の範囲】

【請求項1】計算機により情報を自動的に分類するための方法であって、

階層関係で相互に関連付けされた複数のフォルダの一つにそれぞれ対応し、それぞれ検索すべき一つまたは複数の単語を指定する複数の検索条件を記憶し、

各フォルダに対応して記憶された検索条件と分類すべき文書情報との間の適合度を検出し、

各フォルダと該情報との間の検出された適合度と上記階層関係とに基づいて、該複数のフォルダの内、該文書情報を登録すべき一つまたは複数のフォルダを決定し、

該決定されたする一つまたは複数のフォルダに対応して該文書情報を記憶する情報分類方法。

【請求項2】該検出は、

該文書情報と、各フォルダに対応して記憶された検索条件が指定する単語の各々との間の適合度を検出し、

各フォルダに対応して記憶された検索条件が指定する単語のそれぞれと該文書情報との間に関して検出された適合度の総和を、そのフォルダと該文書情報との間の適合度として決定するステップを含む請求項1記載の文書情報分類方法。

【請求項3】該文書情報と各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々との間の適合度を決定するステップは、その文書情報内でのその単語の重みと、その検索条件内でのその単語の重みとの積をその単語とその文書情報との適合度として決定するステップを有する請求項2記載の文書情報分類方法。

【請求項4】各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々のその文書情報内での重みは、その文書情報内でのその単語の出現回数に比例する値である請求項3記載の文書情報分類方法。

【請求項5】各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々のその検索条件内の重みは、その検索条件内の複数の単語内でのその単語の出現回数に比例する値である請求項3または4記載の文書情報分類方法。

【請求項6】該決定は、

該文書情報と各フォルダとの間に対して検出された適合度により、該文書情報に適合する一つまたは複数のフォルダを検出し、

該検出により複数のフォルダが検出されたときには、それらの検出された複数のフォルダ内の該文書情報を対応させる一つまたは複数のフォルダを、該検出された複数のフォルダの間の、上記階層関係内での相対的位置関係に依存して選択する請求項1記載の文書情報分類方法。

【請求項7】該選択は、該文書情報に適合すると検出された該複数のフォルダの中に、相対的に上下関係にある一群のフォルダが含まれているときには、該一群のフォルダを代表する一つのフォルダを該文書情報を対応させるフォルダとして選択するステップを有する請求項6記載の文書情報分類方法。

2

【請求項8】該選択は、該検出された複数のフォルダの中に、該一群のフォルダと相対的に上下関係にはない他の一群のフォルダが含まれているときには、該他の一群のフォルダを代表する一つのフォルダを該文書情報を対応させる他のフォルダとして検出するステップをさらに有する請求項7記載の文書情報分類方法。

【請求項9】該一群のフォルダを代表する一つのフォルダは、該一群のフォルダの内の最下層に位置するフォルダである請求項7記載の文書情報分類方法。

【請求項10】計算機により文書情報を自動的に分類するための方法であって、

階層関係で相互に関連付けされた複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、

各フォルダに対応して記憶された検索条件と予め定めて判断基準とに基づいて、分類すべき文書情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数を決定し、

決定されたフォルダに対応して該文書情報を記憶し、

複数の分類すべき文書情報の各々に対して上記決定および記憶を行ない、

各フォルダに対応して記憶された複数の文書情報が、そのフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、

いずれか一つのフォルダが該所定の条件を満たしたとき、その一つのフォルダに対応して記憶された複数の文書情報とそのフォルダに対応して記憶された検索条件を再構成するステップを有する文書情報分類方法。

【請求項11】該所定の条件は、該一つのフォルダの登録された文書情報の総数が所定値を越えているということである請求項10記載の文書情報分類方法。

【請求項12】該再構成するステップは、

そのフォルダに対応して記憶された検索条件を、その検索条件が指定する複数の単語の一部をそれぞれ指定する複数の新たな検索条件に分割し、

該一つのフォルダに登録された複数の文書情報を複数の文書情報群に分割し、

該一つのフォルダを新たな複数のフォルダで置換し、

該複数の新たなフォルダの各々に対応して、該新たな複数の検索条件の一つで指定される一つまたは複数の単語と、該複数の文書情報の分割で得られた一部の文書情報を記憶するステップを有する請求項10記載の文書情報分類方法。

【請求項13】該複数の文書情報を分割するステップは、

該一つのフォルダに対応して記憶された検索情報を分割して得られた複数の新たな検索条件に適合する文書情報からなる複数の文書情報部分群に分割するステップから

なる請求項１２記載の文書情報分類方法。

【請求項１４】該再構成するステップは、

該一つのフォルダに登録された複数の文書情報の一部とそのフォルダに対応して記憶された検索条件が指定する複数の単語の一部とを選択し、

該一のフォルダの下位の階層に、少なくとも一つの新たなフォルダを配置し、

該新たなフォルダに対応して、該選択された一部の単語と該選択された一部の文書情報を記憶するステップを有する請求項１０記載の文書情報分類方法。

【請求項１５】該複数の文書情報の一部と該一部の単語を選択するステップは、

該一つのフォルダに対応して該複数の文書情報を、該一つのフォルダに対応して記憶された単語群の一部で検索可能であるが、該単語群のうちの他の一部の単語では検索不可能な一部の文書情報と、該一部の単語および該他の一部の単語のいずれでも検索可能な他の文書情報とに分離し、

該分離で得られた一部の文書情報およびその分離に使用した該一部の単語を選択するステップを有する請求項１４記載の文書情報分類方法。

【請求項１６】該再構成するステップは、

新たなフォルダを生成し、

該一つのフォルダと他の一のフォルダに重複して登録された複数の文書情報を該新なフォルダに対応して登録し、該一のフォルダと該他の一のフォルダに対する、該重複する複数の文書の登録を削除し、

該一つのフォルダに対応して記憶された単語群と該他の一つのフォルダに対応して記憶された他の単語群の内、該重複する複数の文書を検索するための単語群を、該新たなフォルダに対応して記憶するステップを有する請求項１０記載の文書情報分類方法。

【請求項１７】該新たなフォルダは、該一つのフォルダと該他の一のフォルダと同じ階層に配置される請求項１６記載の文書情報分類方法。

【請求項１８】該文書情報を対応させるための一つまたは複数のフォルダの決定は、

各フォルダに対応して記憶された検索条件が指定する単語群に基づいて、該分類すべき文書情報と各フォルダとの間の適合度を検出し、

各フォルダと該文書情報との間の検出された適合度と上記階層関係とに基づいて、該文書情報を登録するためのフォルダとして、該複数のフォルダの一つまたは複数を決定するステップを有する請求項１０記載の文書情報分類方法。

【請求項１９】計算機により文書情報を自動的に分類するための方法であって、

階層関係で相互に関連付けされた複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、

各フォルダに対応して記憶された検索条件と予め定めて判断基準とに基づいて、分類すべき文書情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数を決定し、

決定されたフォルダに対応して該文書情報を記憶し、

複数の分類すべき文書情報の各々に対して上記決定および記憶を行ない、

該複数のフォルダの内の一部の複数のフォルダに対応して記憶された複数の文書情報が、該複数のフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、

いずれかの一部の複数のフォルダが該所定の条件を満たしたとき、該一部の複数のフォルダに対応して記憶された複数の文書情報と、該一部の複数のフォルダに対応して記憶された複数の検索条件を再構成するステップを有する文書情報分類方法。

【請求項２０】該所定の条件は、該一部の複数のフォルダが、上位側のフォルダとそのフォルダの下位側のフォルダに関する条件を含み、

該再構成は、

該下位側のフォルダに対応して登録された文書情報群と該上位側のフォルダに対応して登録された文書情報群とを、該上位側のフォルダと該下位側のフォルダに対して配分し直し、

この配分し直しの後に、該下位側のフォルダに対応して登録された新たな文書情報群と該上位側のフォルダに登録された新たな文書情報群とに基づいて、該下位側のフォルダに対応して登録された文書情報群と該上位側のフォルダに対応して登録された文書情報群とを該上位側のフォルダと該下位側のフォルダに対して配分し直すステップを有する請求項１９記載の文書情報分類方法。

【請求項２１】該所定の条件は、該下位側のフォルダに対応して登録された文書情報の数と該上位側のフォルダに対応して登録された文書情報の数の相対的大きさに関する条件である請求項１９記載の文書情報分類方法。

【請求項２２】該条件は、該下位側のフォルダに対応して登録された文書情報の数が、該上位側のフォルダに対応して登録された文書情報の数より少ないことである請求項２１記載の文書情報分類方法。

【請求項２３】該文書情報を対応させるための一つまたは複数のフォルダの決定は、

各フォルダに対応して記憶された検索条件とに基づいて、分類すべき文書情報と各フォルダとの間の適合度を検出し、

各フォルダと該文書情報との間の検出された適合度と上記階層関係とに基づいて、該文書情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数を決定するステップを有する請求項１９記載の文書情報分類方法。

【請求項２４】データベースを保持す記憶装置と、該計

5

算機からユーザが指定した文書情報を選択的に検索する計算機とを有すす計算機システムにおいて、
階層関係で相互に関連付けされた、ユーザが指定した複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、
そのデータベースに新規に登録される文書情報があるか否かを監視し、
新規に登録された文書情報があるときには、その文書情報と各検索条件との適合度を判別し、
各検索条件と該文書情報との適合度と該階層関係とに基づいて、該文書情報を対応させる一つまたは複数のフォルダを決定する文書情報収集方法。
【請求項２５】該複数のフォルダの各々が、そのフォルダに対応して記憶された複数の文書情報に関連するフォルダの再構成に関する条件を満たすか否かを判別し、
いずれか一つのフォルダが該再構成の条件を満たすとき、新たなフォルダを生成し、
その一つのフォルダに対応して記憶された複数の文書情報の少なくとも一部を検索するための新たな検索条件を生成し、
該新たな検索条件と該一部の文書情報を該新たなフォルダに対応して記憶するステップをさらに有する請求項２４記載の文書情報収集方法。
【請求項２６】いずれかのフォルダが、そのフォルダに対応して記憶された複数の文書情報を複数の新たなフォルダに分割するための分割条件を満たすか否かを判別し、
いずれか一つのフォルダが該分割条件を満たすとき、複数の新たなフォルダを生成し、
その一つのフォルダに対応して記憶された複数の文書情報および検索条件とから、該複数の文書情報を複数群に分割するための複数の検索条件を決定し、
該分割により得られた複数群の文書情報の内の一つの群の文書情報と、該分割により得られた複数群の検索文書情報の内の一つの群の文書情報とを、該新たな複数のフォルダの一つに対応して記憶するステップをさらに有する請求項２４記載の文書情報収集方法。
【請求項２７】ある一組みの複数のフォルダが、それらに重複して対応して記憶された複数の文書情報を分離して記憶するための条件を満たすか否かを判別し、
その一組みのフォルダが、該分離条件を満たすとき、新たなフォルダを生成し、
それらのフォルダに重複して記憶された複数の文書情報を検索するための検索条件を、該一組みのフォルダのいずれかにそれぞれ対応して記憶された検索条件に基づいて生成し、
該新たなフォルダに対応して、該生成された検索条件を記憶し、
該重複する複数の文書情報を、該新たなフォルダに対応するように記憶し直すステップをさらに有する請求項２４記載の文書情報収集方法。

6

【請求項２８】ユーザに提供すべき文書情報を含むデータベースを記憶する手段を有する第１の計算機と、
該第１の計算機と交信して、該データベース内の文書情報を検索するための第２の計算機と、
該第２の計算機に接続された、ユーザが操作可能な端末とを有し、
該端末は、相互に階層関係で関係づけられた、ユーザが指定した複数のフォルダの名称と、それぞれのフォルダに対応してユーザが指定した複数の単語を含む複数の検索条件とを該第２の計算機に送付する手段を有し、
該第２の計算機は、
該送付された複数のフォルダの名称と検索条件を記憶する手段と、
そのデータベースに新規に登録される文書情報があるか否かを該第１の計算機と交信して検出する手段と、
新規に登録された文書情報があるときには、その文書情報と該複数の検索条件との適合度を判別し、各検索条件と該文書情報との適合度と該階層関係とに基づいて、該文書情報を対応させる一つまたは複数のフォルダを決定し、決定されたフォルダの各々に対応して該文書情報とその名称を記憶する手段と、
該端末からの要求に応答して、該複数のフォルダの名称とそれぞれのフォルダに対応して記憶された複数の文書情報の名称とを該端末に送付する手段とを有し、
該端末は、該送付された複数のフォルダの名称を有する複数のフォルダを、該階層関係が識別可能な態様で表示し、該送付された複数の文書情報の名称を、それぞれの文書情報が対応するフォルダに対応して表示する手段をさらに有する文書情報収集システム。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、計算機ネットワークを介して、自動的に情報を収集、分類、整理する情報収集システムに関する。
【０００２】
【従来の技術】計算機ネットワークの整備は急速に進んでおり、オンライン情報検索サービス、ネットニュースからの情報収集、電子メールや電子掲示板を利用した質疑応答といった、いわゆる情報のブロードキャッチが行なえる環境が整いつつある。
【０００３】これらの最新情報の有用性は認識されているものの、次のような点が問題となり、有効に利用されていない。
【０００４】（１）情報源によって利用法が異なり、複数の情報源から情報収集する操作が煩雑である。
【０００５】（２）検索式を論理式で入力しなければならない。所望の情報を得るための適切な検索式を記述するのは難しい。

7

【０００６】（３）収集した情報の分類と整理に手間と時間がかかる。

【０００７】「２１世紀の情報化社会」（日経バイト１９９１年１１月３２０ページ～３３１ページ）に記載されている広域情報サーバWAISは、（１）の問題点をプロトコルを共通化（NISO Z39.50を拡張）し、さらに情報源への接続と検索を自動化することにより解決し、（２）の問題点を関連性フィードバックにより解決した。関連性フィードバックは次のような検索条件の精練手法である。ユーザが検索したい内容を記述すると、それを検索条件としてWAIS はその内容に合う情報を検索し提示する。ユーザがその中から欲しかった情報を選ぶと、WAISはユーザが選んだ情報を検索条件にフィードバックし、検索条件を改善する。この関連性フィードバックを用いた情報検索により、ユーザは検索式を記述することなく所望の情報を検索できるようになった。

【０００８】（３）の問題点を解決するために、様々な文書の自動分類システムが考案されている。

【０００９】たとえば、特開平１－１８８９３４の文書分類システムは、標本文書群を調べることにより、各分野におけるキーワードの出現頻度情報を得て、入力された文書からキーワードを抽出して、分野毎に点数を計算し、最高得点の分野へ分類する。

【００１０】特開昭６３－２１４８３２の通知文書処理システムは、通知文書の書式を解析し、通信文中に出現する単語の重みを分類カテゴリー別に付加し、その総和を求め、最大となるカテゴリーを選ぶことにより分類する。

【００１１】

【発明が解決しようとする課題】WAISは、上記（１）、（２）の問題点は解決したが、収集した情報の分類、整理に関しては配慮していない。

【００１２】階層的に情報を分類整理することが望まれるが、従来の方法では、これに適していなかった。

【００１３】また、（３）を解決する従来の自動分類システムにおいては、分類する分類体系をあらかじめ確立しておく必要があった。さらに、各分野を特徴付けるキーワード群やキーワード群の出現頻度などをあらかじめ与えるか、または求めるかする必要があった。

【００１４】しかし、あらかじめ適切な汎用的分類体系を設けるのは困難である。分類体系が適切でないと、ある分類に多くの情報が集中することがある。ある分類の情報量が多くなり過ぎると、ユーザは収集した情報の全容を把握しにくくなる。

【００１５】また、最先端の分野では多くの人に認められる分類体系や専門用語が確定していないことが多く、しかも頻繁に変更される。最先端の分野に関する文書を従来の自動分類システムで適切に分類するのは難しい。

【００１６】本発明の第１の目的は、階層的に情報を分類整理するのに適した文書情報分類方法、それを使用した文書情報収集方法およびシステムを提供することにある。

8

【００１７】本発明の第２の目的は、収集した文書情報の集まり具合から、分類体系と分類に用いる検索条件の改良を自動的に行なう文書情報収集方法およびシステムを提供することにある。

【００１８】

【課題を解決するための手段】本発明による第１の文書情報分類方法は、階層関係で相互に関連付けされた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群に基づいて、分類すべき情報と各フォルダとの間の適合度を検出し、各フォルダと該情報との間の検出された適合度と上記階層関係とに基づいて、該情報が対応するフォルダとして、該複数のフォルダの一つまたは複数を決定し、該決定されたする一つのフォルダまたは複数のフォルダの各々に対応して該情報を記憶するステップを有する。

【００１９】本発明による第２の文書情報分類方法は、階層関係で相互に関連付けされた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群と予め定めて判断基準とに基づいて、分類すべき情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数を決定し、決定されたフォルダに対応して該情報を記憶し、複数の分類すべき情報の各々に対して上記決定および記憶を行ない、各フォルダに対応して記憶された複数の情報が、そのフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、いずれか一つのフォルダが該所定の条件を満たしたとき、その一つのフォルダに対応して記憶された複数の情報とそのフォルダに対応して記憶された一群の検索条件を再構成するステップを有する。

【００２０】本発明による第３の文書情報分類方法は、階層関係で相互に関連付けされた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群と予め定めて判断基準とに基づいて、分類すべき情報を対応させるためのフォルダとして、該複数のフォルダの一つまたは複数を決定し、決定されたフォルダに対応して該情報を記憶し、複数の分類すべき情報の各々に対して上記決定および記憶を行ない、該複数のフォルダの内の一部の複数のフォルダに対応して記憶された複数の情報が、該複数のフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、いずれかの一部の複数のフォルダが該所定の条件を満たしたとき、該一部の複数のフォルダに対応して記憶された複数の情報と、該一部の複数のフォルダに対応して記憶された一群の検索条件を再構成するステップを有する。

【００２１】

9

【作用】本発明による第１の文書情報分類方法では、各フォルダに対応して記憶された検索条件と検索対象文書情報との適合度と、複数の検索条件の階層構造とを考慮して、検索対象文書情報を対応させるフォルダを決定するので、ユーザが記述した検索条件群を階層構造をなす分類体系であると見做して収集した文書情報を分類できる。

【００２２】本発明による第２の文書情報分類方法では、各フォルダに対応して記憶された文書情報に依存して、各フォルダの分割など、フォルダの再構成をすることが出来る。したがって、検索により得られた文書情報の集まり具合に応じて、自動的に分類体系を変更できる。

【００２３】本発明による第３の文書情報分類方法では、複数のフォルダにまたがるフォルダの再構成をすることが出来る。

【００２４】

【実施例】以下本発明の１実施例について説明する。

【００２５】本実施例の文書情報収集システムが対象とするのは、オンライン文書情報検索サービス、電子メール、電子掲示板などを介して電子的に得ることができる、それぞれユーザにとって意味のある内容を一群の文字で表した情報である。以下このような情報を文書情報とよぶ。

【００２６】これらのサービスは、それぞれ様々な企業や団体により運営されている。以後これらのサービスを情報源と呼ぶ。各情報源が提供する文書情報は、一般に、多岐に亘るので、複数の分野に分けてユーザに提示される。これらの分野をドメインと呼ぶ。ドメインにおいて提供される個々の情報を文書と呼ぶ。文書が検索条件に適合したときに格納する検索結果格納領域をフォルダと呼ぶ。

【００２７】図１に本実施例の文書収集システムと本実施例の文書収集システムが文書収集する外部の情報源とからなるシステム構成例を示す。本実施例の文書収集システムは文書収集クライアント５００と文書収集サーバ５１０とからなる。

【００２８】文書収集クライアント５００はネットワーク上に複数存在して同時に文書収集サーバ５１０にアクセスすることができる。

【００２９】文書収集クライアント５００のメモリ５２２上の文書収集クライアントシステム５０１は、ユーザが、収集した文書を格納するフォルダを作成したり、どのような文書を収集するかを表す検索条件を各フォルダに登録したり、フォルダに収集された文書を見たりするためのグラフィカル・ユーザ・インタフェースを提供する。

【００３０】文書収集サーバ５１０のメモリ５２３上の文書収集サーバシステム１００は、文書収集クライアントシステム５０１からの要求に応じて文書情報を提供する一方で、自動的に、ニュースサーバ５２０や文書サーバ５２１などの外部の情報源から、ユーザが登録した検索条件群に適合する文書を収集し、さらに分類、整理を行う。

10

【００３１】まず、文書収集クライアントシステム５０１について説明する。

【００３２】ユーザが文書収集クライアントシステム５０１を起動すると図３に示すようなインタフェース画面４００をCRT５０２上に表示する。ユーザはこのインタフェース画面４００上で、キーボード５０３、マウス５０４などの入力デバイスを用いて様々な操作を行い、収集した文書を格納するフォルダを作成・消去したり、文書を収集するための検索条件を記述したり、収集結果を見たり、評価したりする。

【００３３】文書収集クライアントシステム５０１が行う処理の流れ図を図７に示す。文書収集クライアントシステム５０１が起動されると、まず文書収集サーバシステムへの接続を行う（ステップ１２０）。次に図３に示すインタフェース画面４００を表示する（ステップ１２１）。

【００３４】この後、イベントループ１２２に入り、ステップ１２３～１２６を繰り返す。即ち、ユーザの操作を受理・解析し（ステップ１２３）、操作に対応するコマンドを文書収集サーバシステム１００に送信し（ステップ１２４）、実行結果を文書収集サーバシステム１００から受信し（ステップ１２５）、その実行結果をインタフェース画面４００に反映させる（ステップ１２６）、という処理を繰り返す。

【００３５】ユーザがメニューから終了を選ぶ操作を行うと、終了コマンドを文書収集クライアントシステム５０１に送信して、イベントループ１２２を抜け、文書収集サーバシステム１００との接続切断処理を行い（ステップ１２７）、終了する。

【００３６】図３に示したインターフェース画面４００の具体例について説明する。この画面は、既にユーザによってフォルダ作成とそのフォルダに収集すべき文書の検索条件登録が行われ、文書収集サーバシステム１００により、ユーザが登録した検索条件群に適合する文書を収集・分類された時点の例である。

【００３７】４０２は、内容を表示中のフォルダの名前である。この例ではuser1 というフォルダの下位ディレクトリであるvoice というフォルダの内容を表示中である。

【００３８】４０３にはフォルダuser1/voiceにユーザが登録した検索条件を表示する。表示されたテキストを直接編集することにより、検索条件の更新を行うことができる。本実施例では、各フォルダに対して記憶された検索条件は、単語（以下ワードと呼ぶ）、あるいは句、あるいは文章など、ユーザが自然語で記述し得るものを列挙したものからなる。

【００３９】４０４にはフォルダuser1/voiceの下位のフォルダの一覧を表示する。各フォルダについて、フォルダ名、フォルダに収集されている文書数、フォルダに対応する検索条件の書き出しを表示している。この例では、user1/voice の下にそれぞれ、recognition とsynthesis の二つの下位フォルダがある。

【００４０】この下位フォルダ一覧の項目をクリックするとクリックされた下位フォルダへ移動することができる。

【００４１】４０５にはフォルダuser1/voiceにすでに収集されている文書の一覧を表示する。

【００４２】各文書について、タイトル、フォルダuser1/voiceの検索条件への適合度、適合した検索条件中のワード、情報源名、ドメイン名などを表示している。

【００４３】この文書一覧の項目をクリックすると、クリックされた文書の内容を見ることができる。文書の内容は４０６に表示される。

【００４４】フォルダの作成・消去はメニュー４０１のFileメニューを使って行う。また、Gotoメニューを使っても、別のフォルダへ移動できる。

【００４５】また、ユーザはメニュー４０１のEdit メニューを使って収集された文書や文書が格納されているフォルダに対して評価を与えることができる。つまり、ユーザが、メニューを用いて、有用／無用な文書である、有用／無用なフォルダである、という評価を与えると、対応するコマンドが文書収集サーバシステム１００に送られる。文書収集サーバシステム１００は、文書やフォルダに対する評価を検索条件に反映させ、次回の文書収集時からよりユーザの意図にあった文書を収集する。

【００４６】サーバ５１０のメモリ上の文書収集サーバシステム１００は、クライアント５００からの要求を処理する一方で、ユーザが作成したフォルダ群と各フォルダに登録した検索条件に基づいて、文書の収集・分類・整理を行う。

【００４７】つまり、文書収集サーバシステム１００は、ニュースサーバ５２０や文書サーバ５２１などの外部の情報源に定期的にアクセスし、前回にアクセスした後で各情報源に蓄積された文書を取得し、ユーザが登録された検索条件群に適合するものを検索する。この際、検索条件中の各ワードの対象文書における出現数を対象文書とその検索条件との適合度とする。適合した検索条件が登録されているフォルダの中から、フォルダの階層構造を考慮して対象文書を分類するフォルダを選び、そのフォルダへ格納する。さらに、多くの文書が蓄積されたフォルダを自動分割するなどの文書の収集状況に応じた文書の整理を行う。

【００４８】なお、文書収集の対象となる外部の情報源は、サーバからアクセス可能な他のネットワーク上に在っても良い。

【００４９】文書の収集・分類・整理についてさらに詳しく説明する前に、まず、ユーザが作成するフォルダと検索条件について図４に示した例で説明する。

【００５０】文書収集サーバシステム１００にユーザ登録を行うと、各ユーザに一つのフォルダが割り当てられる。ユーザは自分に割り当てられたフォルダの下に、自由に、下位フォルダを階層的に作成して、各々のフォルダに対して、そのフォルダにはどのような文書を収集すべきかという検索条件を登録する。

【００５１】図４の例では２人のユーザ user1、 user2が登録されており、それぞれフォルダ５４０、フォルダ５５０が割り当てられている。user1 は、フォルダ５４０の下に階層的にフォルダ５４１－５４４を作成し、各フォルダに検索条件５４５－５４８を登録してある。

【００５２】一方、user2は下位フォルダを作成せず、フォルダ５５０に、興味のある事柄を羅列しただけの検索条件５５１を登録してある。

【００５３】フォルダとフォルダに対応する検索条件は、ユーザが作成、更新するほかに、文書収集サーバシステム１００によっても、文書の収集状況に応じて自動的に作成されたり、更新されたりすることもある。詳しくは後述する。

【００５４】したがって、user2 のように、階層的なフォルダを作成せずに、興味のある事柄を羅列しておくだけでも、収集された文書は自動的に分類・整理される。

【００５５】図２の流れ図に従い、文書収集サーバシステム１００について説明する。

【００５６】文書収集サーバシステム１００は、複数のユーザからの要求にいつでも対応し、同時に定期的に文書の収集を行うために、常にユーザの接続要求があるか、または、一定時間が経過したかを監視している（ステップ１０１）。ユーザが接続要求をした場合には、クライアント要求処理１１０を開始する。一定時間が経過した場合には、文書収集処理１０２を開始する。いずれの場合も、文書収集サーバシステム１００本体の処理は直ちにステップ１０１に戻り、ユーザの接続要求と一定時間経過の監視を続ける。

【００５７】図６にクライアント要求処理１１０の流れとコマンド実行時に用いるデータ構造との対応を示す。

【００５８】クライアント要求処理１１０が開始されるとまず、クライアントからの要求処理を行うための準備として、クライアントとの接続（ステップ１１１）、フォルダテーブルのロード（ステップ１１２）を行う。

【００５９】このあと、クライアントから終了コマンドを受信するまで、クライアント５００から送信されてくるコマンドの解析（ステップ１１３）と実行（ステップ１１４）を繰り返す。

【００６０】終了コマンドを受信して、繰り返しを終了すると、クライアントの切断を行って、クライアント要求処理１１０を終了する。

【００６１】各コマンドの実行時には、必要に応じて各種のテーブルのロード、参照、更新、セーブを行う。

【００６２】たとえば、ユーザがあるフォルダに格納されている文書一覧の表示を要求する操作をすると、文書収集クライアントシステム５０１は対応するコマンドと対象のフォルダ名を送信する。クライアント要求処理１１０はこのコマンドとフォルダ名を受信すると、フォルダテーブルを参照して、そのフォルダに格納されている文書群の情報（各文書のタイトル、適合度、適合した検索条件中のワード、情報源名など）をクライアントへ送信する。

【００６３】図５に示した文書収集処理１０２（図２）の流れと文書収集処理時に用いるデータ構造との対応に従って、文書収集処理について説明する。

【００６４】まず、内部ＤＢ５１１からメモリ上に文書収集用のテーブル（文書番号テーブル３００、フォルダテーブル３１０、ワード・フォルダテーブル３３０、ワード・文書テーブル３５０）をロードする（ステップ１０３）。

【００６５】文書番号テーブル３００は、どのような情報源が利用可能か、各情報源にはどのようなドメインがあるか、それらのドメインにはそれぞれ何番から何番までの文書があり、既に何番までは取得済みであるかという情報を表す。

【００６６】フォルダテーブル３１０は、どのようなフォルダがどのような階層構造を成しているか、各フォルダにはどのような文書が格納されているかを表す。

【００６７】ワード・フォルダテーブル３３０は、各フォルダに対応付けられている検索条件にはどのようなワードが出現するかを表す。

【００６８】ワード・文書テーブル３５０にはどの文書にどのようなワードが出現するかを表している。各テーブルについて詳しくは後述する。

【００６９】次に、各情報源の全ての新文書について、ステップ１０５～１０７を繰り返し実行する。

【００７０】ステップ１０５の新文書取得処理は、各情報源に接続し、文書番号テーブル３００に登録されている文書番号より新しい文書があるかどうか調べ、もしあればその文書を取得する。

【００７１】次に、ステップ１０６の適合度計算が、取得した文書の各フォルダにおける適合度を計算する。まず、取得した文書にどのようなワードが出現するかを表わすフォルダ検索テーブル３７０を作成し、各フォルダにおける適合度を記憶するために適合フォルダテーブル３９０を作成・初期化する。そして、フォルダ検索テーブル３７０とワード・フォルダテーブル３３０とを照合して、適合度をフォルダごとに算出し、適合フォルダテーブル３９０に登録する。適合度計算について詳しくは後述する。

【００７２】次にステップ１０７の文書格納処理が、適合フォルダテーブル３９０に登録された各フォルダにおける適合度と、フォルダテーブル３１０が表わすフォルダ間の階層構造とから文書を格納するフォルダを決定し、その文書をフォルダテーブル３１０とワード・文書テーブル３５０に登録する。文書格納処理について詳しくは後述する。

【００７３】次に、ステップ１０８のフォルダ管理処理が、ワード・文書テーブル３５０が表わす各文書におけるワードの出現頻度分布を用いてフォルダ内の文書を分析し、フォルダの自動分割や統合を行ない、フォルダテーブル３１０とワード・フォルダテーブル３３０とを更新する。詳しくは後述する。

【００７４】以上のステップ１０５～１０７の繰り返し中に更新された文書収集用テーブルを内部ＤＢ５１１へセーブする（ステップ１０９）。

【００７５】ここまでで、一通りの文書収集処理１０２を終了する。

【００７６】以上述べた文書収集処理１０２で用いるデータ構造や処理についてさらに詳しく説明する。

【００７７】文書番号テーブル３００のデータ構造を１３に示す。文書番号テーブル３００は、ハッシュテーブルで、各エントリは図１２に示す文書番号リスト３０２を指している。情報源名とドメイン名を入力とするハッシュ関数の値でエントリを決定する。

【００７８】文書番号リスト３０２は、情報源名へのポインタ３０３、ドメイン名へのポインタ３０４、そのドメインの最古文書の番号３０５、最新文書の番号３０６、文書収集システムが既に収集処理を施した文書の番号３０７、同ハッシュ値の他の文書番号リストへのポインタ３０８の組である。

【００７９】文書番号テーブルは、文書収集を始める際にロードされ、文書を情報源から取得する度に更新される。

【００８０】内部ＤＢには、どのような情報源があるか、どのようなドメインがあるか、どのドメインの文書は何番まで収集処理済みかが記憶されている。まず、内部ＤＢ５１１から、記憶されている情報源名、ドメイン名、既取得文書番号を読み込んで文書番号リスト３０２を作成し、情報源名とドメイン名を入力とするハッシュ関数の値をエントリとして文書番号テーブル３００に登録する。次に、各情報源から各ドメインの最古文書番号、最新文書番号を取得し、文書番号リストに書き込む。このとき文書番号テーブル３００に登録されていないドメインがあれば、これはその情報源において新規に作成されたドメインであるので、既取得文書番号を０として文書番号リストを生成し、文書番号テーブル３００に登録する。

【００８１】たとえば、図１３の文書番号リスト３０２－ａは、internet news という情報源の fj.ai というドメインには、１２３番から１４５番までの文書があ

15

り、そのうち１３０番までは収集処理済みであることを示している。

【００８２】フォルダテーブル３１０のデータ構造を図１５に示す。フォルダテーブル３１０はハッシュテーブルで、各エントリは図１４に示すフォルダリスト３１４を指している。フォルダ名を入力とするハッシュ関数の値でエントリを決定する。

【００８３】フォルダリスト３１４は、フォルダの ID 番号３１５、フォルダ名へのポインタ３１６、上位フォルダを表すフォルダリストへのポインタ３１７、下位フォルダリスト３２１へのポインタ３１８、格納文書リスト３２４へのポインタ３１９、同ハッシュ値の他のフォルダを表すフォルダリストへのポインタ３２０の組である。

【００８４】下位フォルダリスト３２１は、下位フォルダを表すフォルダリストへのポインタ３２２とフォルダリスト３１４で表されるフォルダの他の下位フォルダを表す下位フォルダリストへのポインタ３２３の組である。

【００８５】格納文書リスト３２４は、格納された文書の情報源名へのポインタ３２５、ドメイン名へのポインタ３２６、文書番号３２７、格納文書リスト３２４が表す文書のフォルダリスト３１４が表すフォルダにおける適合度３２８、このフォルダに格納された他の文書を表す格納文書リストへのポインタ３２９の組である。

【００８６】例えば、図１５のフォルダリスト３１４－ａは、フォルダ ID が１００３の voice というフォルダの上位フォルダはフォルダリスト３１４－ｂで表されるフォルダuser1であること、フォルダリスト３１４－ｃで表されるフォルダsynthesisを下位フォルダに持つことと、このフォルダには適合度１３点のinternet news という情報源のfj.ai というドメインの１２０番の文書等が格納されていることとを表している。

【００８７】図１７に示すワード・フォルダテーブル３３０は、ハッシュテーブルで、各エントリは図１６に示すワード・フォルダリスト３３３を指している。ワードを入力とするハッシュ関数の値でエントリを定める。

【００８８】ワード・フォルダリスト３３３は、ワードへのポインタ３３４、フォルダ頻度リスト３４０へのポインタ３３５、同ハッシュ値の他のワード・フォルダリストへのポインタ３３６の組である。フォルダ頻度リスト３４０は、このワードが出現する検索条件に対応するフォルダのフォルダ ID ３４１、検索条件中のワードの出現頻度３４２、他のフォルダ頻度リストへのポインタ３４３の組である。

【００８９】例えば、図１７のフォルダリスト３３３－ａとフォルダ頻度リスト３４０－ａは、言語というワードが、フォルダ ID １００３のフォルダに対応する検索条件中に１回出現することを表し、フォルダリスト３３３－ｂとフォルダ頻度リスト３４０－ｂ、３４０－ｃ

16

は、音声認識というワードが、 ID １００３のフォルダと ID １００４のフォルダのそれぞれに対応する検索条件中に１回づつ出現することを表す。

【００９０】ワード・文書テーブル３５０のデータ構造を図１９に示す。ワード・文書テーブル３５０はハッシュテーブルで、各エントリは図１８に示すワード・文書リスト３５４を指している。ワードを入力とするハッシュ関数の値でエントリを決定する。

【００９１】図１８のワード・文書リスト３５４は、ワードへのポインタ３５５、文書頻度リスト３６０へのポインタ３５６、同ハッシュ値の他のワード・文書リストへのポインタ３５７の組である。文書頻度リスト３６０は、このワードが出現する文書の情報源名へのポインタ３６１、ドメイン名へのポインタ３６２、文書番号３６３、出現頻度３６４、このワードが出現する他の文書頻度リストへのポインタ３６５の組である。

【００９２】例えば、図１９のワード・文書リスト３３４－ａと文書頻度リスト３６０－ａは、言語 というワードが、情報源internet newsのドメインfj.sci.langの５６番の文書に５回出現することを表し、ワード・文書リスト３３４－ｂと文書頻度リスト３６０－ｂ、３６０－ｃは、音声認識というワードが情報源internet newsのドメインfj.ai の１２０番の文書に２回出現し、ドメインfj.sci.langの５６番の文書に２回出現することを表している。

【００９３】フォルダテーブル３１０、ワード・フォルダテーブル３３０、ワード・文書テーブル３５０の内容は、内部ＤＢ５１１に記憶されている。これらのテーブルは文書収集処理１０２が開始されたときやクライアント要求処理１１０が開始された時やコマンド実行時に、必要に応じてメモリ上へロードされ、それぞれの処理を実行中に参照・更新され、終了するときに内部DB５１１にセーブされる。ただし、各テーブルは排他的に更新される。フォルダの作成・削除によるフォルダテーブルの更新、検索条件の更新によるワード・フォルダテーブルの更新は、ただちにセーブされる。

【００９４】例としてワード・フォルダテーブル３５０のロードについて図８に流れ図を示す。ワード・フォルダテーブル３５０のロードは、フォルダテーブル３１０をロードした後で行う。

【００９５】まず、ワード・フォルダテーブル３５０を初期化する（ステップ１６０）。

【００９６】次に、フォルダテーブル３１０に登録されている全てのフォルダについて、フォルダのワード登録（ステップ１６４～１６６）を繰り返す（ステップ１６１）。

【００９７】フォルダのワード登録は、まず、そのフォルダに対応する検索条件を内部ＤＢ５１１からメモリ５２３上に読みこみ、（ステップ１６４）、ワードを抽出する（ステップ１６５）。抽出した各ワードについて図

１６のワード・フォルダリスト３３３を作成し、ワードのハッシュ値を計算して図１７のワード・フォルダテーブル３３０に登録する（ステップ１６６）。

【００９８】全てのフォルダについてワード登録を行うとこの繰り返しを終了し、ワード・フォルダテーブルロード処理１５１を終了する。

【００９９】適合度計算１０６が行なう検索処理について図９に基づいて説明する。

【０１００】この処理は、検索条件群に出現するワードと文書に出現するワードの類似性を調べることにより、取得した文書と各フォルダの適合度を調べる。

【０１０１】ここで本実施例で使用する、検索対象文書といずれかのフォルダとの適合度について説明する。

【０１０２】検索対象文書といずれかのフォルダとの適合度は、いくつかの方法が考えられるが、本実施例では、その文書内のワードのうち、そのフォルダに適合したワード（すなわち、そのフォルダに対応して記憶された検索条件に含まれるワードに一致した、文書内のワード）のそれぞれとそのフォルダとの適合度を求め、それらのワードとそのフォルダとの適合度の総和を求め、この総和をその文書とそのフォルダとの適応度とする。

【０１０３】ここで、そのフォルダに適応したワードとそのフォルダとの適応度もいろいろの方法で求めることが出来るが、本実施例では、より好適なものとして、そのワードのその文書内での重みとそのワードのそのフォルダ内での重みとの積でもってそのワードとそのフォルダの適応度とする。

【０１０４】ここで、そのワードの文書内の重みは、いろいろの方法で検出可能であるが、本実施例では、より好適なものとして、そのワードのその文書内での出現頻度でもって、そのワードのその文書内での重みとする。

【０１０５】さらに、そのワードとそのフォルダとの適応度もいろいろの方法で検出可能であるが、本実施例では、より好適なものとして、そのフォルダに対応して記憶された検索条件内でのそのワードの出現回数を使用する。

【０１０６】従って、本実施例では、そのワードとそのフォルダとの適合度は、そのワードの文書内出現頻度とそのワードのそのフォルダに対応する検索条件内での出現頻度の積でもって表すことが出来、その検索対象文書とそのフォルダとの適応度は、このようにして求めた各ワードの適応度の総和で与えられる。

【０１０７】より具体的には、取得した文書に出現するワードを図２１のフォルダ検索テーブル３７０に登録し、全フォルダに対応する検索条件に出現するワードを登録してあるワード・フォルダテーブル３３０と照合して、フォルダ毎に適合度を集計し、適合度順にフォルダをソートする。

【０１０８】まず、フォルダ検索テーブル３７０の初期化（ステップ１７０）、図２３の適合フォルダテーブル３９０の初期化（ステップ１７１）を行なう。

【０１０９】次に取得文書からワードを抽出し（ステップ１７２）、各ワードをフォルダ検索テーブル３７０に登録する(ステップ１７３)。

【０１１０】フォルダ検索テーブル３７０はハッシュテーブルで、各エントリは、図２０に示すフォルダ検索リスト３７２を指す。ワードを引数とするハッシュ関数の値でエントリを決定する。フォルダ検索リスト３７２は文書中のワードへのポインタ３７３、適合フォルダリスト３８０へのポインタ３７４、文書中の出現頻度３７５、同ハッシュ値の他のフォルダ検索リストへのポインタ３７６の組である。適合フォルダリスト３８０は、ワードが出現する検索条件に対応するフォルダのフォルダ ID ３８１、そのフォルダにおける適合度３８２、他の適合フォルダリストへのポインタ３８３の組である。

【０１１１】たとえば、図２１のフォルダ検索テーブル３３０のフォルダ検索リスト３７２－ａは、言語というワードが検索対象の文書中に２回出現することを表している。まだ検索を実行していないので、適合フォルダリストへのポインタ３７４－ａは ＮＵＬＬ である。同様に、フォルダ検索リスト３７２－ｂ、３７２－ｃはそれぞれ対象文書中に無音時間というワードが３回出現すること、音声認識というワードが５回出現することを表している。

【０１１２】次に、フォルダ検索テーブル３７０とワード・フォルダテーブル３３０を照合し、適合するフォルダがあれば、適合フォルダリストを作成し、フォルダ検索テーブルに登録する（ステップ１７４）。

【０１１３】すなわち、フォルダ検索テーブル３７０に登録されているワードが、ワード・フォルダテーブル３３０にも登録されていれば、フォルダ検索リスト３７２の頻度３７５とワード・フォルダリスト３３３に登録されている各フォルダ頻度リスト３４０の頻度３４２を掛け合わせた値をそのワードの各フォルダにおける適合度として、それぞれに対応する適合フォルダリスト３８０を作成し、フォルダ検索リスト３７２に登録する。

【０１１４】例えば、フォルダ検索テーブル３７０に登録されている言語というワードはワード・フォルダテーブル３３０のフォルダ頻度リスト３４０－ａが示すように、フォルダ ID １００３に対応する検索条件に１回出現している。したがって、言語 というワードの ID １００３のフォルダにおける適合度は２点で、検索実行前にはＮＵＬＬであった適合フォルダリストへのポインタは、図２２に示した検索実行後のフォルダ検索テーブル３７０のように適合フォルダリスト３８０－ａを指す。

【０１１５】同様に音声認識というワードは、ワード・フォルダテーブル３３０のフォルダ頻度リスト３４０－ｂが示すように、ID１００３のフォルダとID１００４のフォルダに対応する検索条件にそれぞれ１回出現している。したがって、音声認識というワードの ID １００３

のフォルダとID １００４のフォルダにおける適合度はそれぞれ５点である。したがって、検索実行後は、図２２のフォルダ検索テーブル３７０に適合フォルダリスト３８０－ｂと３８０－ｃが登録される。しかし、無音時間というワードは、ワード・フォルダテーブルに登録されていない。すなわち適合するフォルダが存在しないということで、検索実行後も適合フォルダリストへのポインタはＮＵＬＬである。

【０１１６】最後にフォルダ毎に、各ワードのフォルダにおける適合度を集計し、適合度が０でないフォルダを適合度の高い順に図２３の適合フォルダテーブル３８０に登録する（ステップ１７５）。適合フォルダテーブルの各エントリは、適合フォルダリスト３８１を指す。適合フォルダリストは、図２０の適合フォルダリストと同じデータ構造であるが、各ワードのフォルダにおける適合度の合計を適合度とし、他の適合フォルダリストへのポインタ３８３は使用しない。

【０１１７】以上の適合度計算によって作成された適合フォルダテーブル３９０とフォルダの階層構造が登録されているフォルダテーブル３１０を用いて、文書格納処理１０９が、文書を格納すべきフォルダを選んで格納する。文書格納処理１０９の流れを図１０に示す。

【０１１８】この文書格納処理は、大きく分けて２段階（ステップ１８０、１８１）からなる。

【０１１９】まず、ステップ１８０では、適合フォルダテーブル３９０に登録された各フォルダにおける適合度とフォルダテーブル３１０に登録されたフォルダの階層構造から、対象文書をどのフォルダに格納すべきかを決定し、対象文書を格納した文書としてフォルダテーブル３１０に登録する。

【０１２０】図２４に格納フォルダの決定方法の説明図を示す。本実施例ではフォルダの階層構造の各枝で適合したフォルダの中で最も下位のフォルダに格納する。 AからH までのフォルダがあり、それぞれ図に示す適合度であった場合には、A—B—D—G という枝の適合したフォルダの中で最も下位のD、同様に枝A—B—E 中のE、枝 A—C—F—H 中のH に格納する。この方法は、下位のフォルダは上位のフォルダの検索条件を継承していると考え、検索条件をより詳しく記述しているフォルダに格納するものである。

【０１２１】次に、ステップ１８１では、対象文書に出現するワードをワード・文書テーブル３５０に登録する。

【０１２２】すなわち、文書に出現する全ワードについて図１８のワード・文書リストを作成し、ワード・文書テーブル３５０に登録する。

【０１２３】この処理により、ワード・文書テーブル３５０には格納された全文書について、各文書にどのようなワードが出現するかが記録される。このワード・文書テーブル３５０は、次に述べるフォルダ管理処理でフォルダ内の文書を分析するのに用いる。

【０１２４】フォルダ管理処理１０８について説明する。

【０１２５】階層構造を成すフォルダに対応する検索条件を分類体系とみなして文書の収集、分類を続けると、文書が特定のフォルダに集中して、フォルダ内の文書数がユーザが把握しきれないほど増えることがある。また、文書が複数のフォルダに重複して格納されることが多くなり、無駄が生じることもある。

【０１２６】これらの現象は、ユーザが適切に分類体系を構成していなかった場合や世間の情勢や研究動向が変化し、分類体系が合わなくなった場合に起きる。

【０１２７】フォルダ管理処理１０８は、各フォルダへの文書の集まり具合を分析することによって、これらの現象を検知し、フォルダの階層構造やフォルダに対応する検索条件を改良する。これにより、フォルダ内の文書数をユーザが把握できる程度の数に抑さえたり、文書が複数のフォルダに不必要に重複して格納されないようにしたりし、フォルダの階層構造を実情にあった体系に維持する。

【０１２８】フォルダ管理処理の流れを図１１に示す。

【０１２９】各フォルダに対してステップ２０１～ステップ２０５を繰り返し行なう（ステップ２００）。

【０１３０】まず、フォルダに格納された文書数を監視する（ステップ２０１）。フォルダにあらかじめ与えた数以上の文書が格納されていれば、そのフォルダ内の文書を統計的手法を用いて分析する（ステップ２０２）。異なった性質のものが混ざり合っている対象の中で、類似している個体を集めてグループに分類する手法はクラスタ分析として知られており、たとえば、「多変量解析ハンドブック」（現代数学社１９８６年）に記載されている。ステップ２０２はクラスタ分析の手法を用いて、フォルダ内の文書に出現するワードの頻度に基づき文書を再分類する。再分類した文書の集合をクラスタと呼ぶ。

【０１３１】次にクラスタ間の関係を分析し、クラスタの階層構造を決定する（ステップ２０３）。ここで行なうクラスタ間の関係解析については後述する。クラスタに対応してフォルダと検索条件を生成し、クラスタの階層構造に対応して階層的にフォルダを作成する（ステップ２０４）。次にワード・文書テーブル３５０から、生成した各フォルダ内の文書に共通して高頻度に出現するワードを抽出し、フォルダに対応する検索条件に加える（ステップ２０５）。これにより、検索条件を精練することができる。

【０１３２】ここまでの処理を各フォルダに施したら、各フォルダに格納された文書群を分析し、フォルダの再構成、すなわち、フォルダの統合、階層構造の変更を行なう（ステップ２０６）。ここで行なう分析については後述する。

【0133】図26－31を使ってステップ206で行なうクラスタ間関係の分析方法を説明する。

【0134】図25のように、ワード群w1とワード群w2からなる検索条件があり、この検索条件に対応するフォルダ450に文書群dが格納されているとする。このとき、このフォルダ内の文書について、ワード・文書テーブルから得られるデータを統計的に分析してえられる、ワードと文書の関係のパタンを図26の451、図26の455、図28の458に示す。

【0135】図26は、文書群dがワード群w1が出現する文書群d1とワード群w2が出現する文書群d2の二つの独立したクラスタに分類される場合である。この場合、ワード群w1からなる検索条件とワード群w2からなる検索条件を生成し、それぞれに対応するフォルダ453、454と両者の上位のフォルダ452を設け、図26に示す階層構造にする。

【0136】図27は、ワード群1のみが出現する文書群d1とワード群w1とワード群w2出現する文書群d2の二つのクラスタに分類される場合である。ワード群w2が出現する文書群にはワード群w1も出現している。そこで、ワード群w1からなる検索条件とワード群w2からなる検索条件を生成し、それぞれに対応するフォルダ456と457を設け、図27に示す階層構造にする。

【0137】図28は、ワード群w1のみが出現する文書群d1とワード群w2のみが出現する文書群d3とワード群w1とワード群w2の両方が出現する文書群d2の3つのクラスタに分類される場合である。この場合、ワード群w1のみからなる検索条件とワード群w2のみからなる検索条件とワード群w1かつワード群w2なる検索条件を生成し、それぞれに対応するフォルダ456、457、458とこれらの上位のフォルダ455を設け、図28のような階層構造にする。

【0138】同じ図25－30を使ってステップ205で行なうフォルダ間関係の分析方法を説明する。

【0139】図27のような階層構造のフォルダがあるときに、フォルダ453とフォルダ454に重複して格納される文書が増えたとすると、ワードと文書の関係のパタンが451のパタンから457かまたは458のパタンに変化したと考えられる。文書が重複して格納されていることは、フォルダテーブル310から検知できる。そこで、フォルダ453、454に格納されている文書について、ワード・文書テーブル350から得られる各文書におけるワードの出現頻度に基づきクラスタ間関係分析と同様の統計的分析を行ってワードと文書の分布のパタンを調べ、パタンに応じてフォルダと検索条件を再構成する。

【0140】また、図27のような階層構造があるときに、フォルダ457に格納される文書の中でフォルダ456に適合しない文書の割合があらかじめ与えられた割合を越えるようになった場合、検索条件の上下関係が変化したことを意味する。このことは、フォルダテーブル310に登録されているフォルダ457に格納された文書のフォルダ456への適合度を調べることにより検知できる。フォルダ456とフォルダ457とに格納されている文書に対しクラスタ分析を行ない、フォルダを再構成する。

【0141】

【発明の効果】本発明によれば以下の効果が得られる。

【0142】（1）ユーザが記述した検索条件群に適合する情報を複数の情報源から収集し、検索条件群の階層構造を分類体系と見做して収集した情報を分類できる。

【0143】（2）各検索条件に対応した検索結果格納領域への情報の集まり具合に応じて、分類体系を変更することができる。

【0144】その結果、ある検索結果格納領域に格納される情報量をユーザがその全体を把握できる程度の数に抑さえることができる。あるいは情勢の変化に応じ、適切な分類体系を維持できる。

【図面の簡単な説明】

【図1】本発明の一実施例のシステム構成図である。

【図2】本実施例の文書収集サーバシステムの流れ図である。

【図3】本実施例のインタフェース画面の例である。

【図4】本実施例のフォルダとフォルダに対応する検索条件の例である。

【図5】本実施例の文書収集処理の流れ図である。

【図6】本実施例のクライアント要求処理の流れ図である。

【図7】本実施例の文書収集クライアントシステムの流れ図である。

【図8】本実施例のワード・フォルダテーブルロード処理の流れ図である。

【図9】本実施例の適合度計算の流れ図である。

【図10】本実施例の文書格納処理の流れ図である。

【図11】本実施例のフォルダ管理処理の流れ図である。

【図12】本実施例の文書番号リストのデータ構造である。

【図13】本実施例の文書番号テーブルのデータ構造である。

【図14】本実施例のフォルダリストと下位フォルダリストと格納文書リストのデータ構造である。

【図15】本実施例のフォルダテーブルのデータ構造である。

【図16】本実施例のワード・フォルダリストとフォルダ頻度リストのデータ構造である。

【図17】本実施例のワード・フォルダテーブルのデータ構造である。

【図18】本実施例のワード・文書リストと文書頻度リ

ストのデータ構造である。

【図１９】本実施例のワード・文書テーブルのデータ構造である。

【図２０】本実施例のフォルダ検索リストと適合フォルダリストのデータ構造である。

【図２１】本実施例のフォルダ検索テーブルのデータ構造である。

【図２２】本実施例の検索処理後のフォルダ検索テーブルの例である。

【図２３】本実施例の適合フォルダテーブルのデータ構造である。

【図２４】本実施例の文書を格納するフォルダの決定方法の説明図である。

【図２５】本実施例のフォルダ内文書分析を行なうフォルダの説明図である。

【図２６】本実施例のフォルダ内文書分析結果のワードと文書の第１の分布パタンそれに従って生成されるフォルダの説明図である。

【図２７】本実施例のフォルダ内文書分析結果のワードと文書の第２の分布パタンそれに従って生成されるフォルダの説明図である。

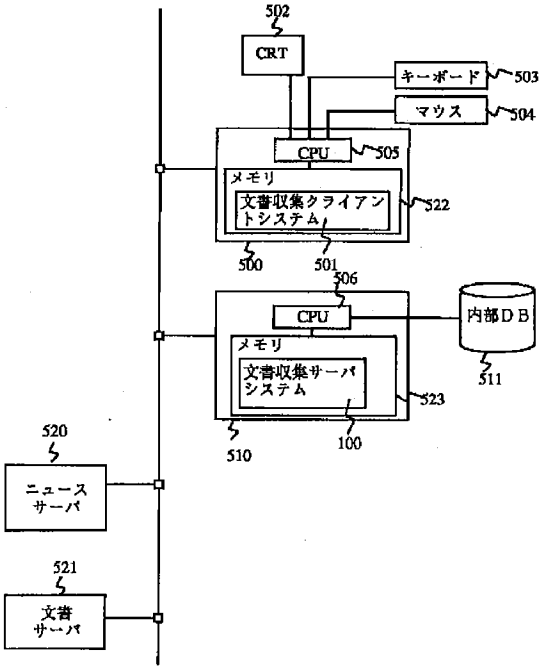
【図２８】本実施例のフォルダ内文書分析結果のワードと文書の第３の分布パタンそれに従って生成されるフォルダの説明図である。

【符号の説明】

１００…文書収集サーバシステム、１０２…文書収集処理、１０６…適合度計算、１０７…文書格納処理、１０８…フォルダ管理処理、１１０…クライアント要求処理。

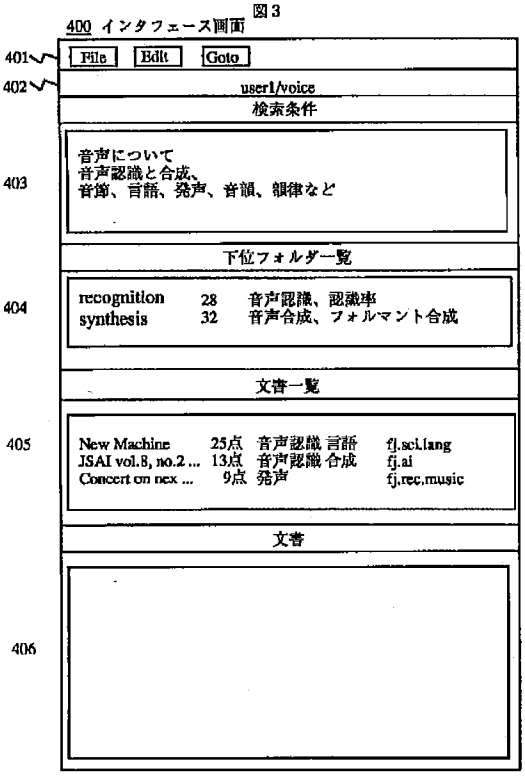
【図１】

図１



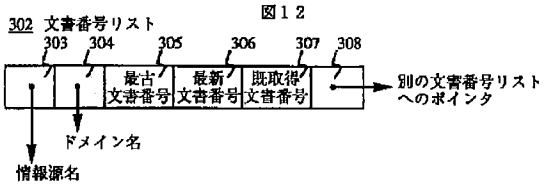
【図３】

図３



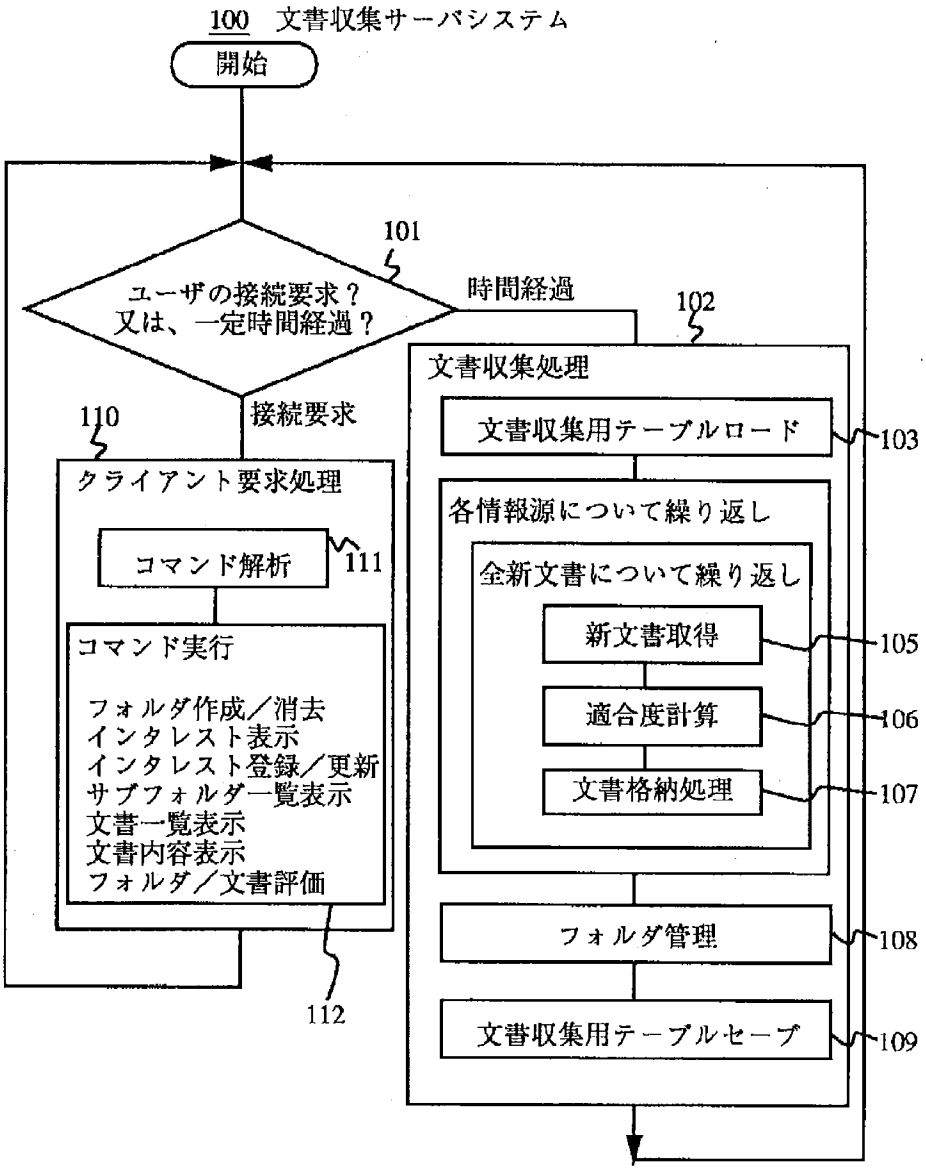
【図１２】

図１２



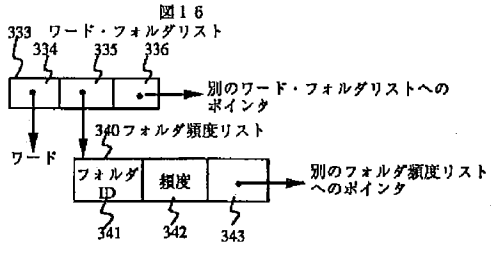
【図２】

図２



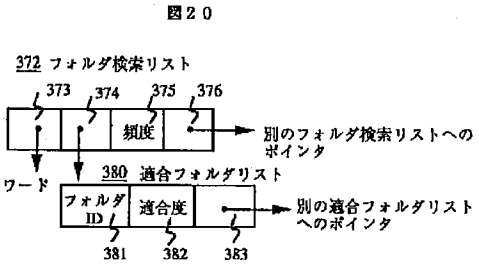
【図１６】

図１６



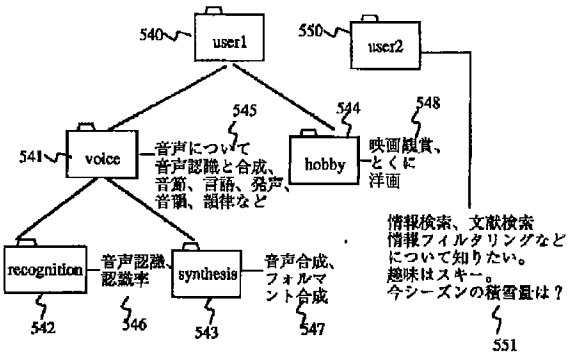
【図２０】

図２０



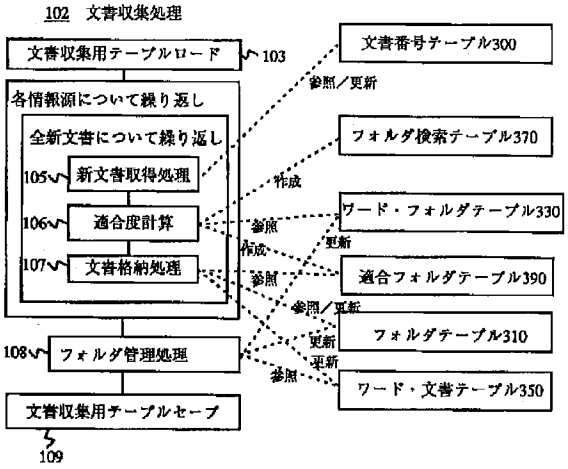
【図４】

図４



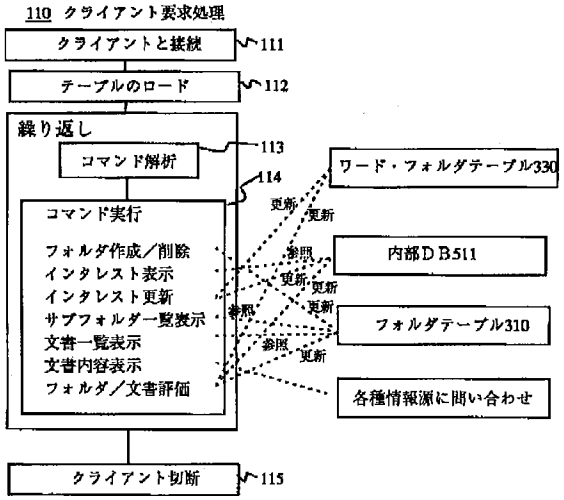
【図５】

図５



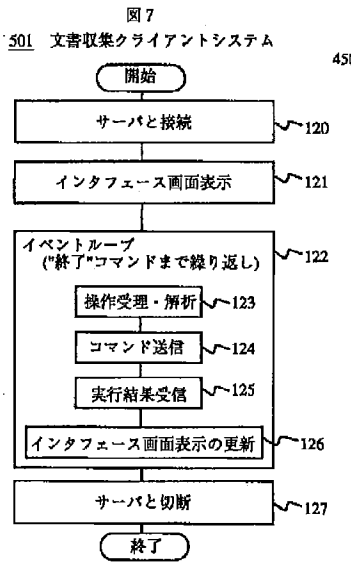
【図６】

図６



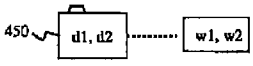
【図７】

図７



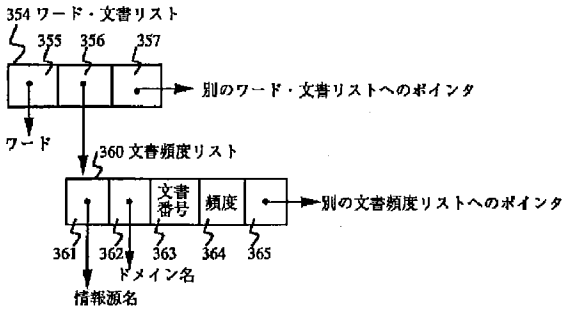
【図２５】

図２５



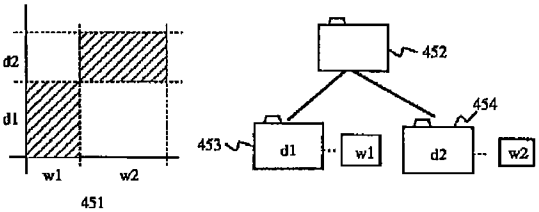
【図１８】

図１８



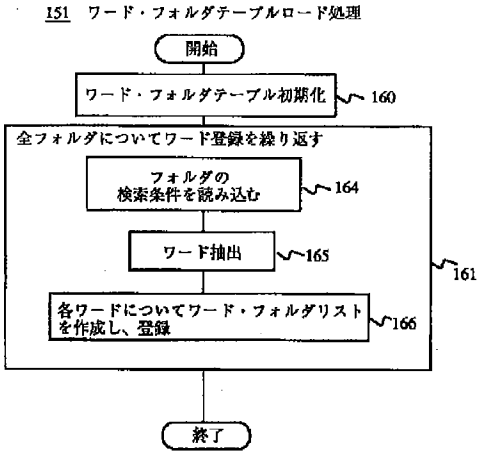
【図２６】

図２６



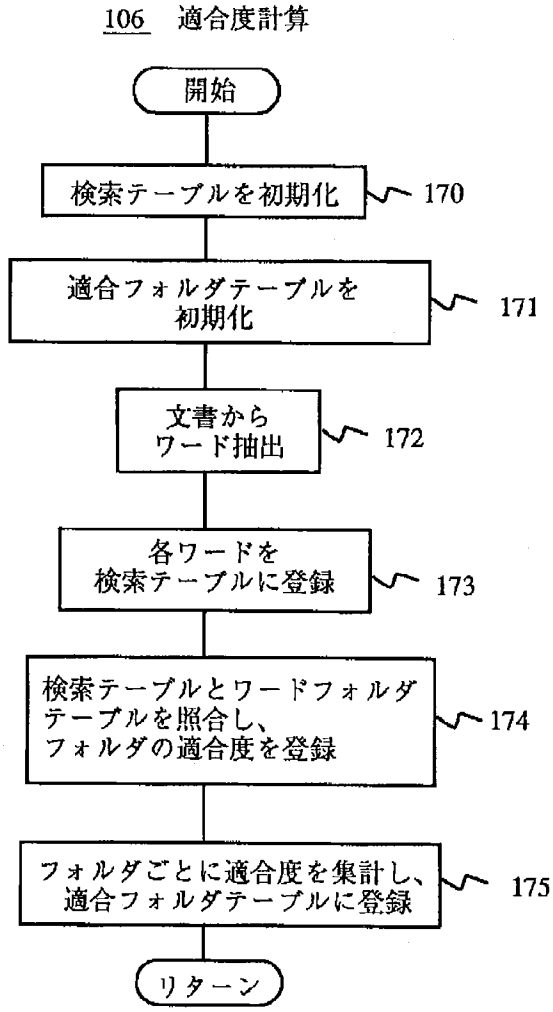
【図８】

図８



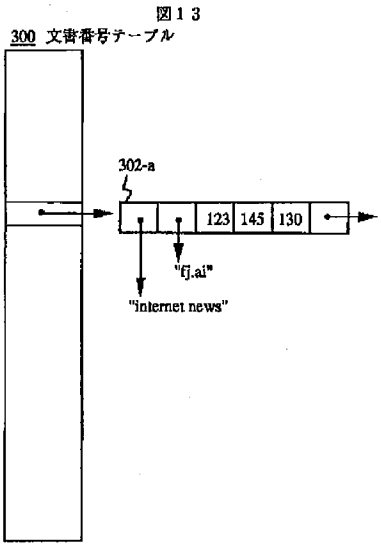
【図９】

図９



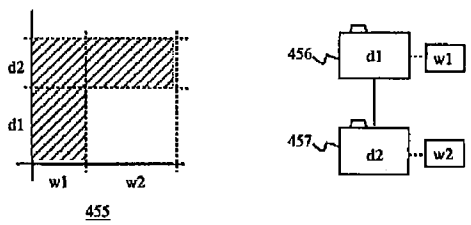
【図１３】

図１３



【図２７】

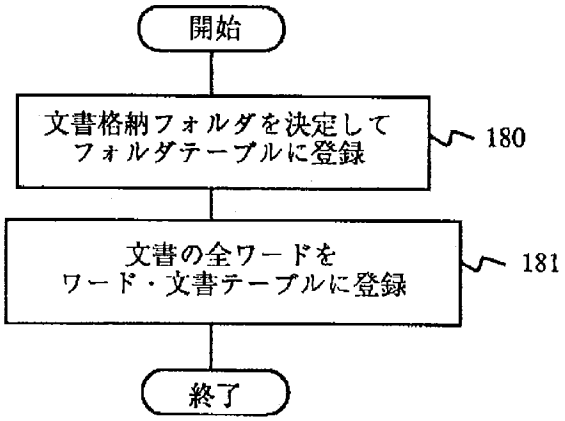
図２７



【図１０】

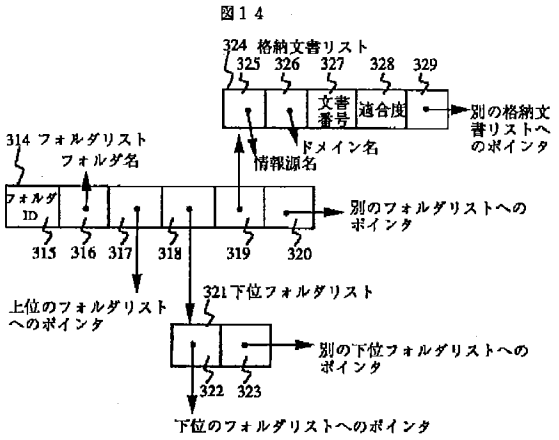
図１０

107　文書格納処理



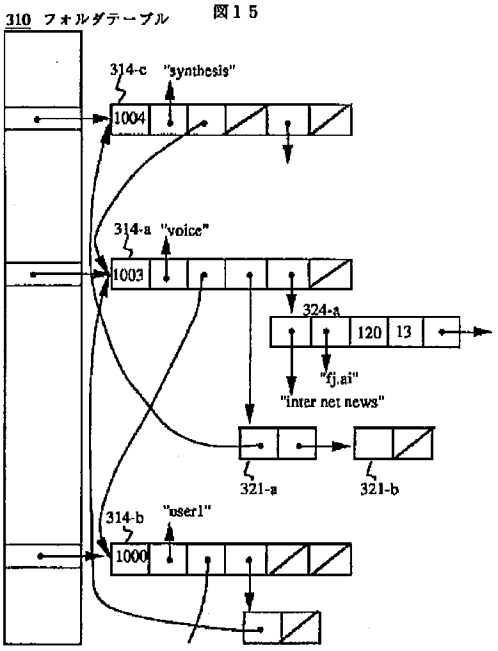
【図１４】

図１４



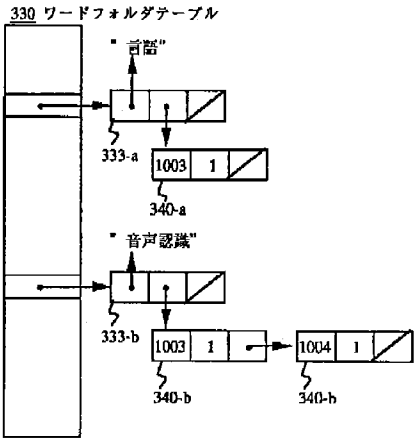
【図１５】

図１５



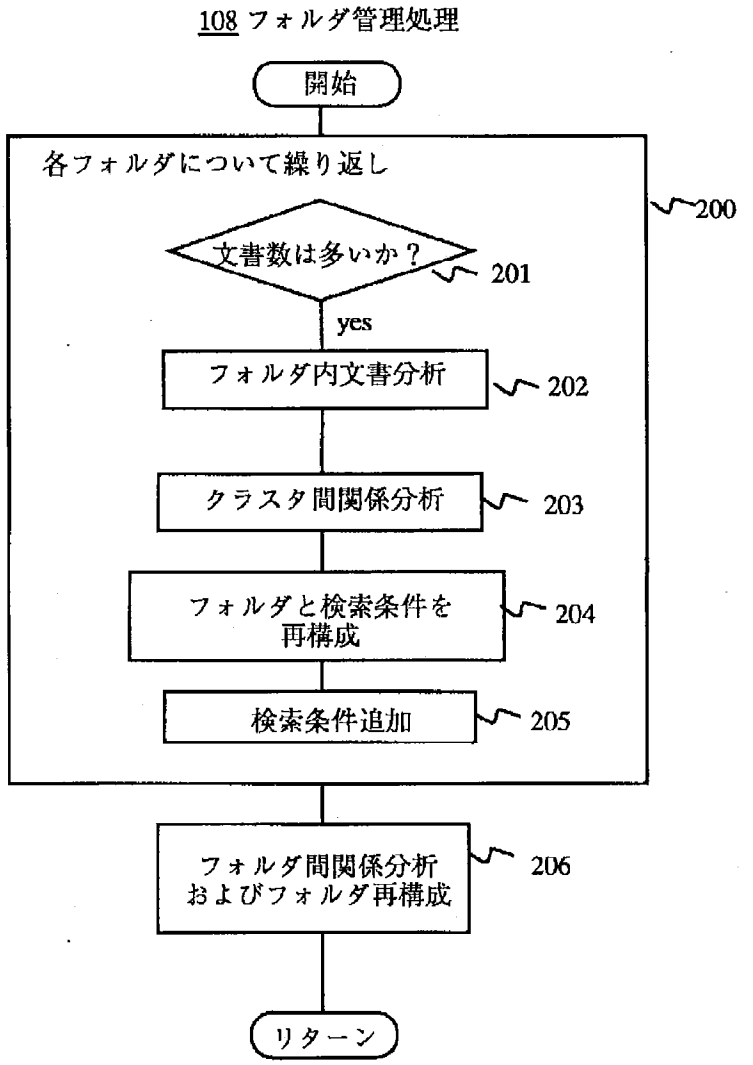
【図１７】

図１７



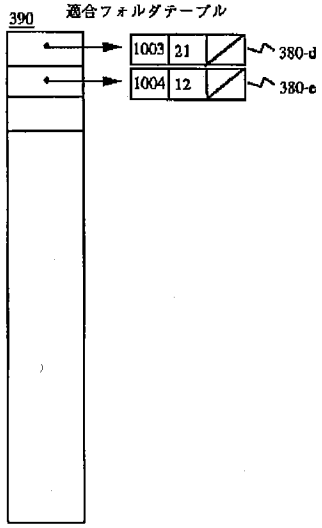
【図１１】

図１１



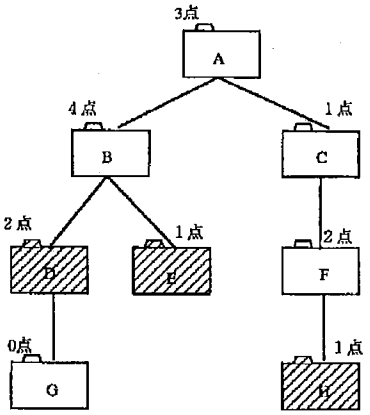
【図２３】

図２３



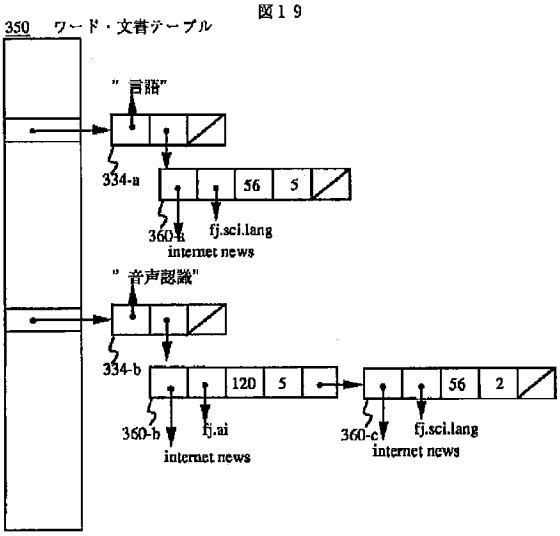
【図２４】

図２４



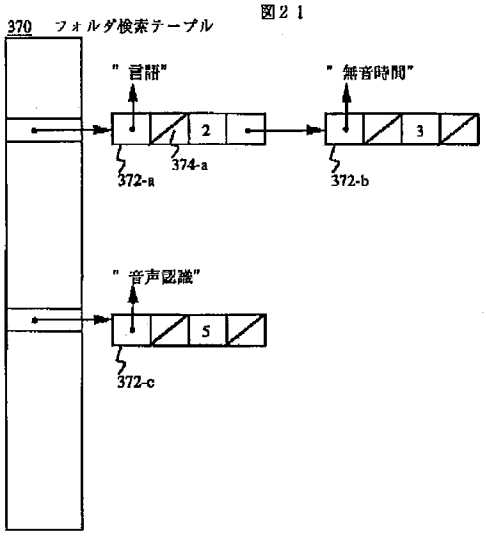
【図１９】

図１９



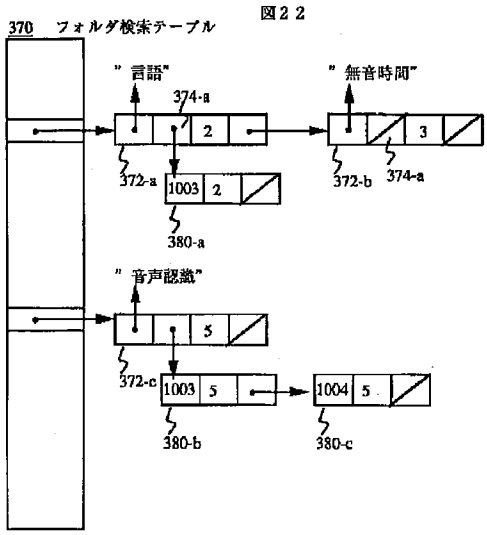
【図２１】

図２１



【図２２】

図２２



【図２８】

図２８